Document IDs
or
International Standard Book Numbers for the Electronic Age

Brewster Kahle
Thinking Machines
5/9/90

A document identifier, if implemented well, will allow a user to know if she has two references to the same document and provide an address to aid in retrieval. This brief paper will suggest and implementation of Document ID's (doc-id) for electronic publications that can be used with the Z39.50 standard. Further this paper will try to list a set of likely scenarios that will show how these ID's can be used.

The simplest use of a doc-id is to receive it from a server as a response to a search, and then retrieve the document by passing it back to the server.

The rough goals of the implementation of the document-id structure are to be:

1) easy to create unique ID's for documents (without a central authority),
2) possible to retrieve the document using the ID (serve as an address),
3) allow users of the ID's to know the copyright intent of the publisher,
4) and be terse.

The design I will suggest in this paper has a long form and a short form. I will describe the long form first and the show how it can be shortened.

There are several fields to a doc-id, each an arbitrary length string except the last field:

```
Original-server
Original-database
Original-local-ID
Distributor-server              ;;optional
Distributor-database            ;;optional
Distributor-local-ID            ;;optional
Copyright-disposition
```

Roughly, the "original" server/db/local-id triple is the original publisher of the document. This can be used to figure out if two documents are identical even if they have been retrieved through different distributors. The distributor server/db/local-id triple is a legitimate distributor of the document so that the original source does not have to be queried each time a user wants the document. The copyright-disposition field has one of three values: copy-without-restriction, all-rights-reserved, and distribution-restrictions-apply. More details are below.

When the original server gives out a doc-id it does not have to supply a distributor triple since it would be redundant with the origin triple. In fact, the original server only has to give out the local-id and the copyright-disposition since the server and the database is known to the client. Short form from the original server is:

Original-local-id and
Copyright-disposition.
The short form from a distributor is:
Original-server,
Original-database,
Original-local-ID,
Distributor-local-ID, and
Copyright-disposition.
The client will fill in the rest of the origin slots as needed.

Doc-ID's will often be sent in a shortened form, but even if it isn't it should be many fewer than 100 characters long.

SERVER

The Original-server and Distributor-server are unique identifiers for the servers. The suggested way to make it a unique identifier is to use a name (or an address) of the server machine on a network. In other words, if a machine used its IP address, then it is guaranteed to be unique since Internet addresses are maintained by an organization for this purpose. Similarly, the server's phone number would also be unique.

Further, in many circumstances, this address can be used to direct retrieval requests. Thus, the doc-id would serve as an address of the document. This would be true if the origin (or client) were on the network that the address is valid for. If the origin were not, then other contact addresses can be retrieved from a directory of servers by using the address as the name.

A server, in its directory description, can specify its server name explicitly; or it can default to its Internet address if any, then phone number (including country code) if any, then X.25 address if any, in that order. In practice, there will be some limit on the length of the address, but each implementation should obey some minimum (80? 255?).

Even in the case of redistribution, is it not legitimate to change the original-server/db/local-id slot. Similarly, the copyright-disposition should not be changed. Changing these slots amounts to claiming ownership and may be legally wrong.

DATABASE

The original-database and distributor-database are copies of the field used in a Z39.50 request. These fields are specific to the server. An example database is "DowQuest" to the DowJones server. Lists of available databases within a server is presumably available through the explain service offered by Z39.50.

LOCAL-ID

The original-local-id and the distributor-local-id are unique identifiers within that database. Some databases may name them in a human readable way, such as "NYTimes 3/14/89 #34", or as just a number. A database on the

server should be able to take a local-id and know what document it refers to. Of course, it is possible for the original document to be deleted in which case, the user will get an error, but real publishers generally try to keep back copies of old periodicals.

Do we need versions, and what would they do if we had them

COPYRIGHT-DISPOSITION

Copyright-disposition is an 8 bit field that has only 3 values defined. This field is the least thought through; I dont understand the legal implications to say if this will hold up in court. This definition is trying to satisfy a number of known scenarios.

Value:	Meaning:
0	copy-without-restriction
1	all-rights-reserved
2	distribution-restrictions-apply

This field is set by the original-server and should never be changed. Question: should we have another value for your-eyes-only which means to not even redistribute the document-id? The assumption is that document-id's are free to be given out; access restrictions are done when retrieving the document's contents.

Copy-without-restriction means that the document may be reproduced in part or in entirety without contacting the original server. This does not mean that the material is not copyrighted. The text of the document should contain copyright information in it. A user that retrieved such a document could serve it on a local system if she wanted to.

All-rights-reserved means that the document should not be given out to other users (though the document-id and headline can be). The distributor should be contacted to get a copy. If no distributor is specified, then the original-server should be contacted.

Distribution-restrictions-apply is a general case to cover unknown future situations. The meaning of this value is dependent on the server and database. Therefore a publisher could define it to mean that you can distribute the document within your site, but not externally. This negotiation of the meaning is not handled within this protocol, rather it is defined in the description of the server or some other way. DowVision, for instance, will probably have this value on its documents since it can be distributed within the site from one machine. Thus DowVision might send its documents to that select machine with this bit set, but with the distributor slot empty. Then the select machine would fill the distributor slot but not change the copyright-disposition slot (it is not legitimate to change that slot in any circumstances).

LIKELY SCENARIOS USING DOCUMENT IDs

If a client asks a question of a netnews (or some other uncopyrighted) server the response would come back with the Original-local-id and copyright-disposition set to copy-without-restriction. If the client redistributed this document, then that machine would fill in the

original-server and original-database slot with the correct values. Also it would generate a distributor-local-id for the document sometime before it is sent out to a requesting machine. It is optional for this machine to fill in the distributor-server and distributor-database since the requesting machine can fill in these slots.

If a client wants to save enough information about the document to look at it again, then the client would want to save the full document id (with the server slots filled in if they are not already), the headline, the best segment, and the score.

If a major publisher is shipping some documents to a redistibutor, then it would fill in the original-local-id slot and the copyright-disposition slots. The redistributor would then fill in the original-server and original-database slots (if needed) and the distributor-local-id slot before redistributing it. This would guarantee that the distributor was asked for copies of the document rather than the original server. If the copyright-disposition is set to all-rights-reserved or distribution-restrictions-apply, then no other machines will overwrite the distributor slots. If the copyright-disposition is set to copy-without-restriction, then other machines could overwrite the distributor slot. A redistributor is not free to change the copyright-disposition to make it the sole redistributor since the copyright-disposition is a property that is assigned by the original-server.

If a server created a document specifically for a client (on the fly, say), then a local-id will only be valid for a short amount of time. There is no way, in this scheme, to specify when this ID will turn invalid. This is a restriction, but should not impede most uses.

Other information about a document might be included in a response from a server such as the headline, the score (how appropriate the server thought the document answered the question), and the best segment of the document. These fields are not included in the doc-id. This is somewhat of an arbitrary decision, but terseness argues for the minimum in the doc-id. The original field and the copyright fields appear useful even if the headline is not available.

PROBLEMS:

A useful address for many servers is a telenet or tymnet address. Since the phone numbers vary in local areas, this does not make sense. Should we invent a syntax such as "dow@telenet" and the client machine uses a local telenet number to get in and then knows to type "c dow"?

Should the original and distributor slots be an ascii string? This will make some implementations easier, but it might make other languages difficult to support. Are there international issues in dealing with this problem?

- Justin 577- 6619

✓ Evie 267-5115
7PM Design Court
648 Beacon St 4th floor Beacon street
Kenmore

✓ Donna 445-6999
643-4850

Patrick Bray

Cliffla
Robin Palmer

Evie
Stan Kugell 469-5545
Karl 628-7403
Alain Tues dinner
Robin Palmer

Pamich
Pam. C